# Applying Automated Analysis to Develop a Cost-Effective Measure of Science Teacher Pedagogical Content Knowledge

Molly Stuhlsatz, Chris Wilson, Zoë Buck Bracey, Mark Urban-Lurain, John Merrill, Kevin Haudek

Teacher pedagogical content knowledge (PCK), or the type of teacher knowledge that bridges content knowledge and how to effectively teach the content in classrooms, has been shown to be a significant predictor of both reform-based classroom practice and student achievement in science (Roth et al., 2011). However, despite the importance of this construct, current measures of PCK have limited use, in part due to them being highly time and resource intensive to score and in part because there are no widely accepted PCK frameworks allowing inferences to be made across measures. Although the field has yet to arrive at a consensus model for PCK, such a model is beginning to emerge from efforts such as the PCK Summit, held at BSCS in October 2012, and from a growing body of literature over the last several decades (Berry, Friedrichsen, & Loughran, 2014). This project builds on the prior work of the field to develop a measurement instrument for assessing teacher PCK in science through a video analysis task and automated computer scoring.

Our primary research question for this project is

- How can lexical analysis and machine learning techniques be applied to developing an efficient, valid, and reliable measure of teacher pedagogical content knowledge?

With the following secondary research questions:

- Can we develop automated computer scoring models of teachers' written responses that closely correlate with expert human coding?
- What feedback can we provide from the automated computer scoring that will facilitate quantitative research and evaluation, professional development and teacher education, and teacher self-evaluation?

This paper focuses on the development of a framework for the construct that is being used to develop a measure of teacher PCK.

## Theoretical Framework

### Defining PCK

In October 2012, science education researchers and leaders from around the world came together in Colorado Springs, CO, for a summit on the development and measurement of pedagogical content knowledge. We have taken our broad definition of PCK from the summit as follows: "Personal PCK is the *knowledge* of, *reasoning* behind, and *planning* for teaching in a particular *topic* in a particular *way* for a particular *purpose* to particular *students* for enhanced *student outcomes*" (Gess-Newsome, 2015, p. 36). From this work emerged a complex model of teacher professional model and skill in

which PCK is embedded as one of several dynamic elements. An important lesson to take from this model is the distinction between PCK and PCK&S, which stands for pedagogical content knowledge and skill (Gess-Newsome, 2015). While PCK is a knowledge base, PCK&S is a knowledge base in action in the classroom. It is important to note that in the development of the framework described here we intend to measure PCK only, which does not require the teacher to be able to apply her or his knowledge—this is why we do not need any classroom observation. As Gess-Newsome (2015) note in their description of the PCK Summit model, there is a tension between what teachers know and what they can actually do, and this distinction is important. Using this framework we are measuring only what teachers know, not what they can actually do. In addition to constructing a synthesis model of PCK,

1) PCK exists on a continuum from weak to strong;
2) PCK can be strengthened through teaching experience, professional development, or other;
3) teaching experience does not necessarily result in increased PCK;
4) teachers with strong PCK are better able to improve student learning;
5) PCK can be found in two forms: knowledge and enactment;
6) enactment of PCK is more difficult to assess and may not lend itself to normative judgments;
7) the explicit knowledge form of PCK may be easier to assess.

PCK represents a complex construct that is difficult to define and even more difficult to measure. Current measures of PCK are highly time and resource intensive to score. Our goal with this project was to use automated analysis to develop an assessment instrument that is grounded in the emerging consensus model for PCK, developed over several decades, but efficient enough to be implemented on a large scale. Such an instrument would primarily be used for research purposes, but we also recognize the potential for future applications to be used as a formative assessment tool for teachers interested in improving their practice.

**The Assessment Triangle**

Effective educational assessments, whether for students or teachers, must carefully coordinate and align three key components: cognition, observation, and interpretation. This coordination is illustrated in the assessment triangle (Figure 1) described in the National Academy of Sciences publication *Knowing What Students Know* (National Research Council [NRC], 2001). The cognition portion of the triangle refers to the models of knowledge, skills, and learning within the domain that is being assessed. The observation component describes the tasks learners are asked to perform to demonstrate their knowledge and/or skills, such as answering assessment items or demonstrating proficiency via performance tasks. The interpretation component includes the tools to make sense of the observations, such as scoring guides, rubrics,

and measurement models. These three elements, both alone and combined, are essential in ensuring that the assessment provides meaningful information about student or teacher understanding. That is, the information that an assessment provides should be the primary focus of all instrument development activities.

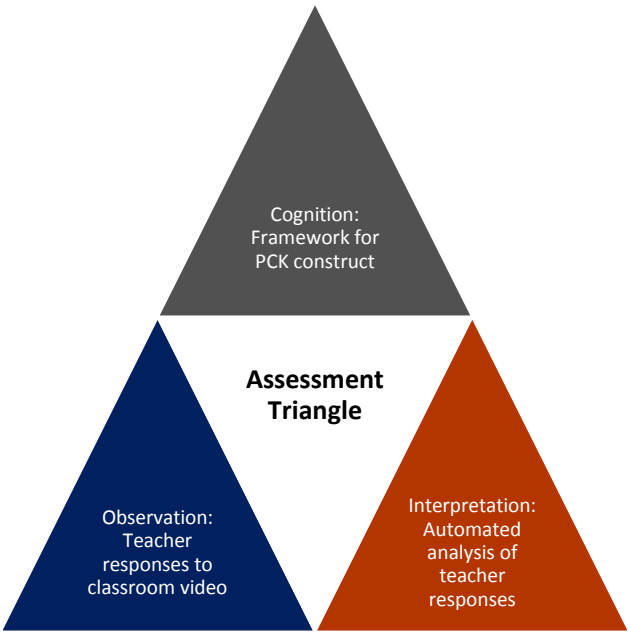


*Figure 1. The Assessment Triangle, adapted from NRC (2001).*

This focus on information provided by assessments is also prominent in test validity theory. The modern, unified view of validity emphasizes the consequential basis of test validity in addition to the more commonly addressed empirical basis. While these two constructs are not always easily separated, the main argument is that tests are not inherently valid but instead are only valid for particular uses or decisions (Messick, 1995). That is, the empirical basis of test validity is not sufficient: We cannot merely run statistical tests, surpass commonly cited cutoffs on various indices, and declare a test valid. Instead, while statistical properties are still required, the modern view of validity requires test scores to be informative and to allow those administering the test to make the decisions the test was designed to inform (Wilson, Roth, Taylor, Landes, & Stuhlsatz, 2012). In this project we bring together the recommendations illustrated in the assessment triangle with the unified view of test validity to develop an instrument that will provide meaningful information.

Our ultimate purpose is to refine the interpretation corner of the triangle to develop effective research tools for measuring PCK, with the secondary goal of formative assessment. This paper focuses on the cognition portion of the triangle, which represents our conceptual framework for PCK. Classroom video analysis tasks provide

the observation portion, and human-scored rubrics provide the interpretation. Through the question development cycle (QDC, Figure 2), the next phase in this project will be to improve on each aspect of the assessment triangle through iterative cycles, refining our rubric (cognition), collecting new data (observation), and eventually replacing human scoring in the interpretation portion with effective automated scoring. The QDC describes the cycle of development created by Urban-Lurain et al. (2013) at Michigan State University to produce a predictive model for automated scoring.

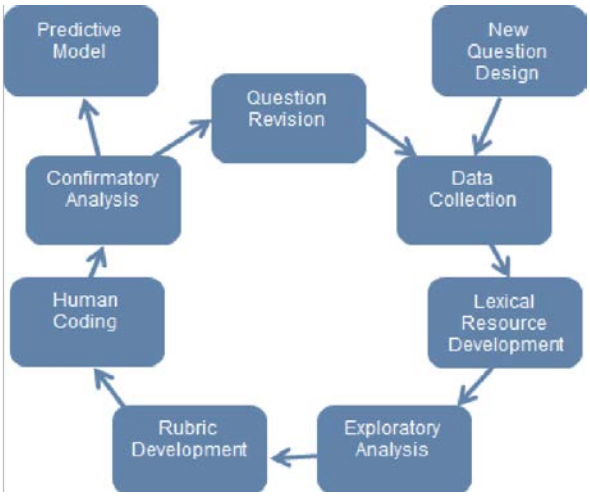


*Figure 2. Question Development Cycle (QDC) (Urban-Lurain et al., 2013).*

The first stage of the QDC is typically to design new questions to measure participant thinking. This is followed by data collection using those questions and lexical resource development using lexical analysis software to extract key terms and scientific concepts from the writing. These terms and concepts are used as variables for exploratory analysis which aids in rubric development. We use both analytic and holistic rubrics for human coding of responses. During confirmatory analysis the lexical resources are used as dependent variables in statistical and machine classification techniques to predict expert human coding of responses. However, in this project we have begun the cycle without going through the exploratory analysis and rubric development stages, with the development of a rubric grounded in existing theory. A large body of work defining PCK has already been done, and we draw on this work to define the construct. Though we began the cycle with a rubric based on the emerging consensus model for PCK, our development process is ongoing, and we will continue to refine the rubric through future iterations of the cycle using data from participants. Thus the cognition corner of the triangle, representing the PCK construct, will become more valid over iterations of the QDC.

## Methods

### Research Design

As described above, the QDC drives the design of our research, which is proceeding in two phases. In Phase 1 of the project we have been working with a large data set of responses to adapt an existing PCK open-response instrument to align it with the literature on PCK, the findings from the PCK Summit (Berry et al., 2014), and the practices of science outlined in the NGSS. That is, we are defining the construct of the assessment and addressing the cognition corner of the assessment triangle as discussed above (rubric development in the QDC).

The existing PCK instrument asks teachers to watch video clips of science classrooms and demonstrate their PCK in written responses where they describe their observations of pedagogical moves and student thinking. The existing data set provides a corpus of responses that has been scored by human coders using a rubric (human scoring). At this point we have not yet developed lexical resources, so we have refined the rubric based on human scoring and identified new video clips based on our refined rubric (question revision). These clips provide a higher level of alignment with the PCK framework we developed based on the emerging consensus model. Next, we will collect a new set of responses from teachers (data collection).

Following new data collection we will move into Phase 2 of the project and continue through the QDC (albeit not always in a linear fashion), refining and iteratively improving all aspects of the assessment, developing lexical resources, and finally moving into predictive scoring models. To do this we will use lexical categories as independent variables in models that predict the human scoring as the dependent variable. Moving through the QDC will ultimately produce an instrument and predictive scoring models that will provide meaningful information to multiple stakeholders that will inform decisions regarding research, evaluation, professional development, and teacher education.

### Cognition: Building a Framework for the PCK Construct

We have broken PCK into six dimensions, described in detail in Table 1. Our six dimensions are rooted in Magnusson's seminal model of the five dimensions of PCK: knowledge about student thinking, knowledge about instructional strategies, knowledge about assessment, knowledge about curriculum, and orientation toward teaching (Magnusson, Krajcik, & Borko, 1999). However, our interpretation of these dimensions has been informed by several more contemporary efforts, including the conclusions of educational leaders from the PCK Summit, the well-established BSCS STeLLA strategies for science teaching (Roth et al., 2011; Taylor, Roth, Wilson, Stuhlsatz, & Tipton, 2016), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), and several factors that research has shown make an impact on equitable learning outcomes in science education.

*Table 1. PCK Framework*

| Dimension | Overview |
|---|---|
| Contextualized analysis of student thinking (CAST) | noticing student thinking re: domain content and/or bringing in knowledge of common patterns of student thinking in the domain |
| Contextualized analysis of instructional coherence (CAIC) | noticing how the teacher does/does not maintain a coherent content storyline |
| Contextualized analysis of constructivist instructional strategies (CACIS) | noticing how the teacher does/does not invoke teaching strategies that allow students to construct their own understanding individually or in groups |
| Contextualized analysis of responsive teaching (CART) | noticing how the teacher does/does not respond to student thinking through formative assessment, "teachable moments," or allowing inquiry to be student led |
| Contextualized analysis of NGSS practices (CANP) | noticing how classroom activity does/does not align with any of the eight NGSS practices (does not need to explicitly mention NGSS) |
| Contextualized analysis of scientific discourse and language in the classroom (CASD) | noticing how science/everyday language is/is not used/scaffolded in the classroom |

We have left the first category of Magnusson's model (student thinking) relatively untouched, as it is considered fundamental to the construct. This has been demonstrated by its inclusion in almost every recent framework for PCK found in the science education literature (e.g., Lee, Brown, Luft, & Roehrig, 2007; Krauss, Brunner, Kunter, Baumert, Blum, Neubrand, & Jordan, 2008; Padilla, Ponce-de-León, Rembado, & Garritz, 2008; Schneider & Plasman, 2011; Brown, Friedrichsen, & Abell, 2013; Gess-Newsome, 2015). This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Student Thinking (CAST). The use of the word *contextualized* in each dimension of our framework is a deliberate attempt to preserve to importance of student and domain context within all aspects of PCK and is explained in more detail in the next section.

The category of instructional strategies is also fundamental in the PCK literature, but because of potential boundary issues we have narrowed the focus to be on instructional strategies that the literature on learning outcomes in science education has shown to be effective over the last several decades—constructivist and social constructivist instructional strategies (Bransford, Brown, & Cocking, 1999). We take constructivism to broadly encompass instruction designed to allow students to build their own knowledge through a combination of strategies including well scaffolded inquiry, questioning that pushes student thinking, and the use of participation structures such as small groups to capitalize on social construction of knowledge. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Constructivist Instructional Strategies (CACIS).

We have also narrowed the category of assessment to focus specifically on formative assessment, which is emphasized across the NGSS, the PCK Summit, and the STeLLA strategies and included in the literature on equity as a productive and authentic

way of measuring student learning and improving teaching (Darder, 1991; Wilson & Sloane, 2008; Solano-Flores, 2008; Moschkovich, 2007; Gipps, 1999). Because our respondents are reacting to short clips of classroom video and are not privy to how the teacher uses assessment longitudinally nor to the design of paper assessments, this category has been altered to better reflect how a teacher can act formatively on student thinking in the moment. We call this "responsive teaching." This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Responsive Teaching (CART).

We have also reframed orientation toward science teaching to focus on those practices of science emphasized in the NGSS, in a dimension we call analysis of NGSS practices. This dimension looks for how respondents are making observations and suggestions related to scientific practices in the classroom such as constructing explanations, communicating ideas, or engaging in arguments. We believe that there is a strong link between a respondent's ability to recognize scientific practice as an important aspect of classroom instruction and their orientation toward science teaching as more than a didactic exercise. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of NGSS Practices (CANP).

We have adjusted the curriculum category to align with what STeLLA research has shown to be fundamental: instructional coherence (Roth et al., 2011). Henze, van Driel, & Verloop (2008) describe the knowledge and beliefs about curriculum to be focused on the purposes of the content in curricular materials—which is essentially staying true to a learning goal that is appropriate for both the topic and the students. In her chapter based on the PCK Summit, "Model of Teacher Professional Knowledge," Gess-Newsome, (2015) defines curricular knowledge as including "the goals of a curriculum ... the role of a scope and sequence, and the ability to assess a curriculum for coherence" (p. 32). Respondents in this case do not have access to the full curriculum, but what they see in the classroom is an enactment of that curriculum, and a coherent curriculum is reflected in instruction that maintains a coherent science content storyline (Roth et al., 2011). Thus we aim to measure the observations and suggestions that respondents make regarding learning goals and the coherence of classroom activity around those goals. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Instructional Coherence (CAIC).

In addition to these dimensions we have added a category on scientific discourse and language. The literature on equity, particularly for students from nondominant cultural and linguistic backgrounds, emphasizes the importance of linguistic supports in the science classroom (Lemke, 1990; Lee & Fradd, 1998; Ash, 2003; Mosqueda, 2010; Lemke, 2001; Shaw, Bunch, & Geaney, 2010; Fradd, Lee, Sutman, & Saxton, 2002). Building off of this, the NGSS (NGSS Lead States, 2013) suggest that when "supported appropriately," providing "rich opportunities and demands for language learning" (p.50) in the science classroom can be very beneficial for English language learners and other students from nondominant linguistic backgrounds. In addition, anecdotal evidence from the STeLLA study indicates the potential of linguistic scaffolds for improving the

effectiveness of science teaching for all students, regardless of linguistic background (Taylor et al., 2016). Thus we seek to measure how respondents are making observations and suggestions related to scaffolding students' development of scientific discourses and vocabulary. This aspect of PCK is incorporated into the dimension of our framework called Contextualized Analysis of Scientific Discourse (CASD).

We use the word *contextualized* to distinguish a simple analysis of content knowledge or teaching strategies from a true analysis of PCK. Shulman's (1986) original conception of PCK was as the *intersection* of content and pedagogy, and thus it is vital that we do not try to measure these elements individually. Schneider and Plasman (2011) warn that PCK should not be measured on individual dimensions of subject matter, pedagogical, and context knowledge. The construct of PCK is a transformation not an integration of these dimensions (Magnusson et al., 1999), and thus they cannot be teased out and measured on their own scale (Gess-Newsome & Lederman, 1999).

Schneider and Plasman (2011) frame PCK as knowledge that is complex but less situated than other types of knowledge that a teacher might draw on in specific classroom situations, allowing it to be measured across domains and across settings—but this does not mean that PCK can be decontextualized from science content and/or student experience. Research has shown that there are no universal heuristics for the classroom that can be applied across contexts to create a "good teacher," and a rubric that claims to measure such a universal, decontextualized construct should be suspect. According to Magnusson, Krajcik, and Borko, who developed one of the seminal models for PCK still used by researchers, PCK "is a teacher's understanding of how to help students understand *specific subject matter*" that includes knowledge about how that subject matter can be "organized, represented, and adapted to *the diverse interests and abilities of learners*, and then presented for instruction" (1999, p. 96, emphasis added). Thus, picking out teacher observations and/or suggestions about instruction is not measuring PCK unless attention is given to subject matter and learner experience. In fact, Brown, Friedrichsen, and Abell (2013) found that even the Magnusson model was not coherent enough to adequately capture the fluid, integrated nature of PCK and recommend a more integrated model. The definition arrived upon at the PCK Summit emphasizes that PCK is "context specific" knowledge—emphasizing that such knowledge is attached to "the teaching of particular topic in a particular way for a particular purpose to particular students" (Gess-Newsome, 2015, p. 36).

The word *contextualized* in our rubric is meant to indicate an analysis of teacher responses that takes into account the importance of how each dimension of PCK is inextricable from the other dimensions and in particular from understanding content and understanding students. We are seeking teacher observations and/or suggestions that are rooted in an understanding of the importance of both domain and of a situated understanding of student thinking and experience. Thus, a teacher who simply lists teaching strategies that could be used in a classroom will not score as highly as a teacher who describes one strategy but considers it deeply and in the appropriate context.

The inclusion of subject matter and student context in each and every dimension of our framework means that we do not have discrete categories for identifying content knowledge or context knowledge. While we recognize that those dimensions are sometimes included in research (e.g., Rowan, Schilling, Ball, Miller, Atkins-Burnett, & Camburn, 2001), they are not included in the Magnusson model nor in most subsequent models (e.g., Lee et al., 2007; Krauss et al., 2008; Padilla et al., 2008; Schneider & Plasman, 2011; Brown et al., 2013; Gess-Newsome, 2015). Thus, we do not consider them to be adequate measures of PCK without being incorporated into more practical pedagogical dimensions.

**Observation: Teacher Responses to Classroom Video**

The assessment task that comprises the observation corner of the assessment triangle is an innovative video-based lesson analysis task (Figure 3). Teachers watch video clips of science lessons across several content areas. These video clips are carefully selected from authentic classroom video to include a range of student activity and teacher pedagogical moves. Teachers respond to a prompt to make analytical comments about the science content, the teaching, and/or the students. In response to the prompt, teachers are given 20 minutes to watch the clip and then provide a written analysis, with their responses varying between a few sentences and three paragraphs.


*Figure 3. Video-based lesson analysis task.*

At this point we have not yet developed lexical resources, so we have refined the rubric based on human scoring and identified new, targeted video clips based on our refined rubric. These clips provide a higher level of alignment with the new PCK framework. We hope that by reducing the length of the video clips, focusing on just one or two categories of the rubric, overall teacher response lengths will be reduced. Shorter teacher responses, which we expect to include only these categories, should allow us to identify lexical resources more easily, complete human coding more quickly, and ultimately refine our computer models efficiently.

**Interpretation: Human and Automated Analysis of Teacher Responses**

We chose to apply a binary scoring scheme to each of the dimensions of the PCK construct in our rubric as a way of increasing inter-rater reliability between human coders and ultimately increasing the likelihood of success with the computer models. We have already completed one round of human coding based on these six dimensions, using teacher responses to the original lesson analysis task. It was clear to us after this first round of scoring that the length of the responses was difficult for both human scoring and machine scoring. Between our two trained human scorers we found that some categories were quite easy to reach agreement, while others were more difficult for the coders to score. Table 2 shows the Cohen's kappas between the two new coders and the expert coder. Both Coder 1 and Coder 2 were more likely to agree with the expert than with each other, and the most difficult coding category was Contextualized Analysis of NGSS Practices (CANP).

*Table 2. Cohen's kappa for inter-rater reliability.*

| | Coder 1 and Expert | Coder 2 and Expert | Coder 1 and Coder 2 |
|---|---|---|---|
| **Contextualized analysis of student thinking (CAST)** | .78 | .67 | .52 |
| **Contextualized analysis of instructional coherence (CAIC)** | .68 | .77 | .55 |
| **Contextualized analysis of constructivist instructional strategies (CACIS)** | .68 | .68 | .56 |
| **Contextualized analysis of responsive teaching (CART)** | .55 | .76 | .60 |
| **Contextualized analysis of NGSS practices (CANP)** | .31 | .37 | .38 |
| **Contextualized analysis of scientific discourse and language in the classroom (CASD)** | .93 | .89 | .85 |

The next step in this project will be to apply human and automated analyses to the new responses gathered through observation—this represents the interpretation corner of the assessment triangle. After we collect a corpus of responses to the new video tasks, we will again go through steps in the QDC to identify lexical resources, do new human scoring, and then use machine learning to see if we can train the computer to score the responses with the same or better reliability as we are able to achieve with human raters. The final product of the QDC is a predictive model that can be used to completely automate the scoring of a new set of responses, predicting how experts would score the responses.

**Conclusion**

The ultimate goal of this project is to develop effective and efficient tools for measuring PCK that are grounded in theory and informed by data. We recognize the challenges of measuring PCK yet see potential in the development of computer-generated scoring strategies using lexical analysis and machine learning as a way to

economically include the construct in large-scale research in the future. We are developing these tools by cycling through the QDC, thereby iteratively improving every corner of the assessment triangle as it applies to this measure.

It is widely acknowledged that new standards like the NGSS will only have an impact on teaching and learning if there are high quality assessments that are closely aligned with the standards (NRC, 2012). While much attention is currently being placed on challenges associated with student assessment, the measurement of teacher-level variables is of equal importance. In order to help teachers develop understandings of these complex standards frameworks and the abilities to integrate science ideas, practices, and crosscutting concepts, we need to be able to measure these understandings and abilities. We expect that findings from this study will provide ample evidence for the efficacy of scoring complex teacher understandings without the need for highly time intensive and expensive human scorers.


This material is based upon work supported by the National Science Foundation under Grant 1437173. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

**References**


Ash, D. (2003). Dialogic inquiry in life science conversations of family groups in a museum. *Journal of Research in Science Teaching, 40*(2), 138-162.

Berry, A., Friedrichsen P., & Loughran, J. (Eds.). (2014). *Re-examining Pedagogical Content Knowledge in Science Education*. New York, NY: Routledge.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Brown, P., Friedrichsen, P., & Abell, S. (2013). The development of prospective secondary biology teachers PCK. *Journal of Science Teacher Education, 24*(1), 133-155.

Darder, A. (1991). *Culture and power in the classroom: A critical foundation for bicultural education.* New York, NY: Greenwood Publishing Group.

Fradd, S. H., Lee, O., Sutman, F. X., & Saxton, M. K. (2002). Materials development promoting science inquiry with English language learners: A case study. *Bilingual Research Journal, 25*(4), 479-501.

Gess-Newsome, J., & Lederman, N. G. (Eds.). (1999). *Examining pedagogical content knowledge: The construct and its implications for science education* (Vol. 6). Norwell, MA: Kluwer Academic Publishers.

Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK: Results of the thinking of the PCK summit. In Berry, A., Friedrichsen P., & Loughran, J. (Eds.). *Re-examining Pedagogical Content Knowledge in Science Education*. New York, NY: Routledge.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355-392.

Henze, I., van Driel, J. H., & Verloop, N. (2008). Development of experienced science teachers' pedagogical content knowledge of models of the solar system and the universe. International Journal of Science Education, 30(10), 1321-1342.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology, 100*(3), 716.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English-language backgrounds. *Educational Researcher, 27*(4), 12-21.

Lee, E., Brown, M. N., Luft, J. A., & Roehrig, G. H. (2007). Assessing beginning secondary science teachers' PCK: Pilot year results. *School Science and Mathematics, 107*(2), 52-60.

Lemke, J. L. (1990). *Talking Science: Language, Learning and Values. Language and Educational Processes*. Westport, CT: Ablex Publishing.

Lemke, J. (2001). Articulating communities: Sociocultural perspectives on science education. *Journal of Research in Science Teaching, 38*(3):296-316.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Examining Pedagogical Content Knowledge: The Construct and Its Implications for Science Education. Science & Technology Education Library.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Moschkovich, J. N. (2007). Beyond words to mathematical content: Assessing English learners in the mathematics classroom. In A. H. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 345-352). New York, NY: Cambridge University Press.

Mosqueda, E. (2010). Compounding inequalities: English proficiency and tracking and their relation to mathematics performance among Latina/o secondary school youth. *Journal of Urban Mathematics Education, 3*(1), 57-81.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Padilla, K., Ponce-de-León, A. M., Rembado, F. M., & Garritz, A. (2008). Undergraduate professors' pedagogical content knowledge: The case of 'amount of substance'. *International Journal of Science Education, 30*(10), 1389-1404.

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., & Camburn, E. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study*. Ann Arbor, MI: Consortium for Policy Research in Education, University of Pennsylvania.

Roth, K. J., Garnier, H., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching, 48*(2), 117-148.

Schneider, R. M., & Plasman, K. (2011). Science Teacher Learning Progressions A Review of Science Teachers' Pedagogical Content Knowledge Development. *Review of Educational Research, 81*(4): 530-565.

Shaw, J. M., Bunch, G. C., & Geaney, E. R. (2010). Analyzing language demands facing English Learners on science performance assessments: Development and use of the SALD framework. *Journal of Research in Science Teaching, 47*(8).

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher, 15*(2), 4-14.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language learning in the testing of English language learners. *Educational Researcher, 37*(4), 189-199.

Taylor, J. A., Roth, K. J., Wilson, C. D., Stuhlsatz, M. A. M., & Tipton, E. (in press). The Effect of an Analysis-of-Practice, Videocase-Based, Teacher Professional Development Program on Elementary Students' Science Achievement. *Journal of Research on Educational Effectiveness*.

Urban-Lurain, M., Prevost, L., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013, October 26). *Using computerized lexical analysis of student writing to support just-in-time teaching in large enrollment STEM courses*. Paper presented at the Frontiers in Education, Oklahoma City.

Wilson, C. D., Roth, K. J., Taylor, J. A., Landes, N. M., & Stuhlsatz, M. A. M. (2012, April). *In search of instructional sensitivity: The measurement problem in large-scale studies of professional development programs.* Paper presented at the National Association for Research in Science Teaching, Indianapolis, IN.

Wilson, M., & Sloane, K. (2008). From principles to practice: An embedded assessment system. In H. Wynne (Ed.), *Student assessment and testing: Vol. 3* (pp. 87-112). Thousand Oaks, CA: Sage.